Computational, Data, and Collaboration Grids

Distributed High-Performance Computing, Large-Scale Data Management, and Collaboration Environments for Science and Engineering

William E. Johnston[*]*, Dennis Gannon*[**]*, and Bill Nitzberg*[***]

[*] National Energy Research Scientific Computing Division, Lawrence Berkeley National Laboratory and Numerical Aerospace Simulation Division, NASA Ames Research Center – wejohnston@lbl.gov

[**] University of Indiana and NAS, NASA Ames – gannon@cs.indiana.edu.

[***] bill@computer.org

Outline

Overall Motivation and Goals

Large-scale science and engineering is typically done through the interaction of

- **people,**
- **heterogeneous computing resources,**
- **multiple information systems, and**
- **instruments,**

all of which are *geographically and organizationally dispersed*.

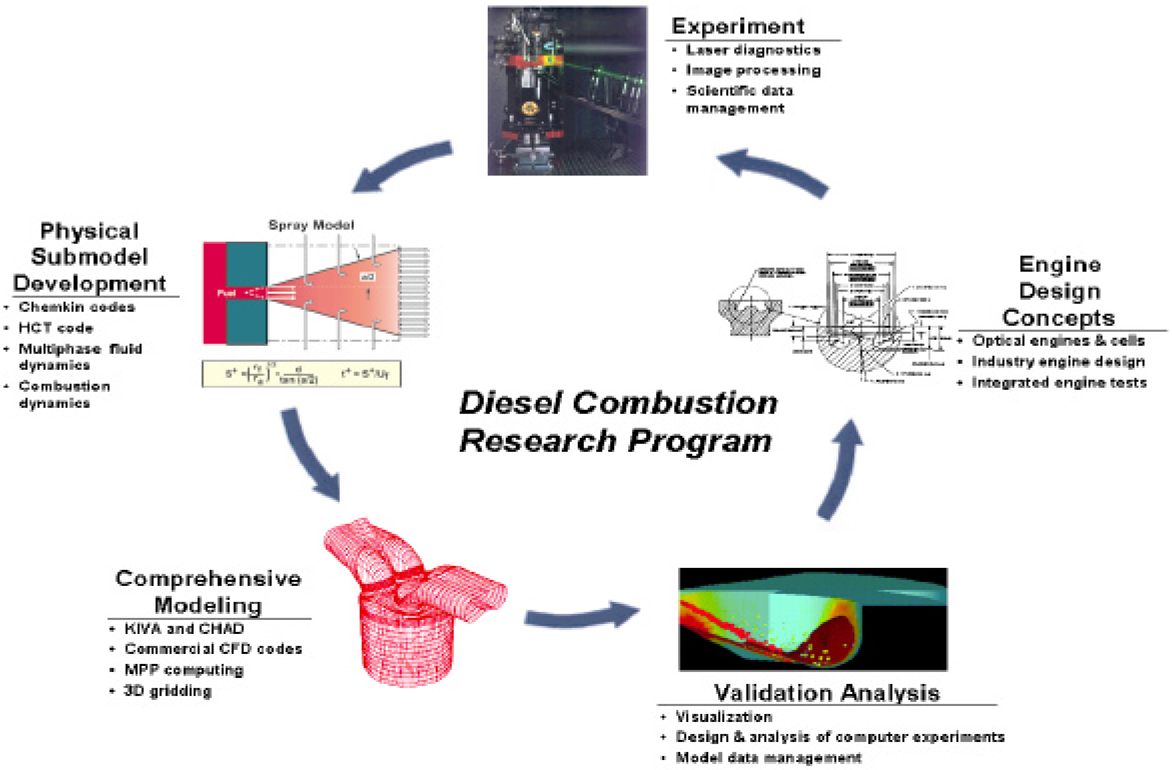
The overall motivation for "Grids" ([3],[4]) is to *enable the routine interactions* of these resources to facilitate this type of large-scale science and engineering.

Applications

***Several types of science and engineering scenarios* are driving the development and deployment of Grids in DOE and NASA:**

- ***Scientific data analysis and computational modeling with a world-wide scope of participants* – e.g. High Energy Physics data analysis**

- ***Large-scale, multi-institutional engineering design* and multi-disciplinary science – e.g., design of next generation diesel engines, next generation space shuttle, etc.**



Through computer and information technologies, the DoE Diesel Combustion Collaboratory [5] aims to reduce geographical barriers to information transfer and increase the utilization of computational models and visualization.

Applications

- ***Real-time data analysis for on-line instruments*****, especially those that are unique national resources – e.g. LBNL’s and ANL’s synchrotron light sources, PNNL’s gigahertz NMR machines, etc.**

- ***Coupling of laboratory instrument experiments*** **and computational models to support, e.g., experiment and computational steering**

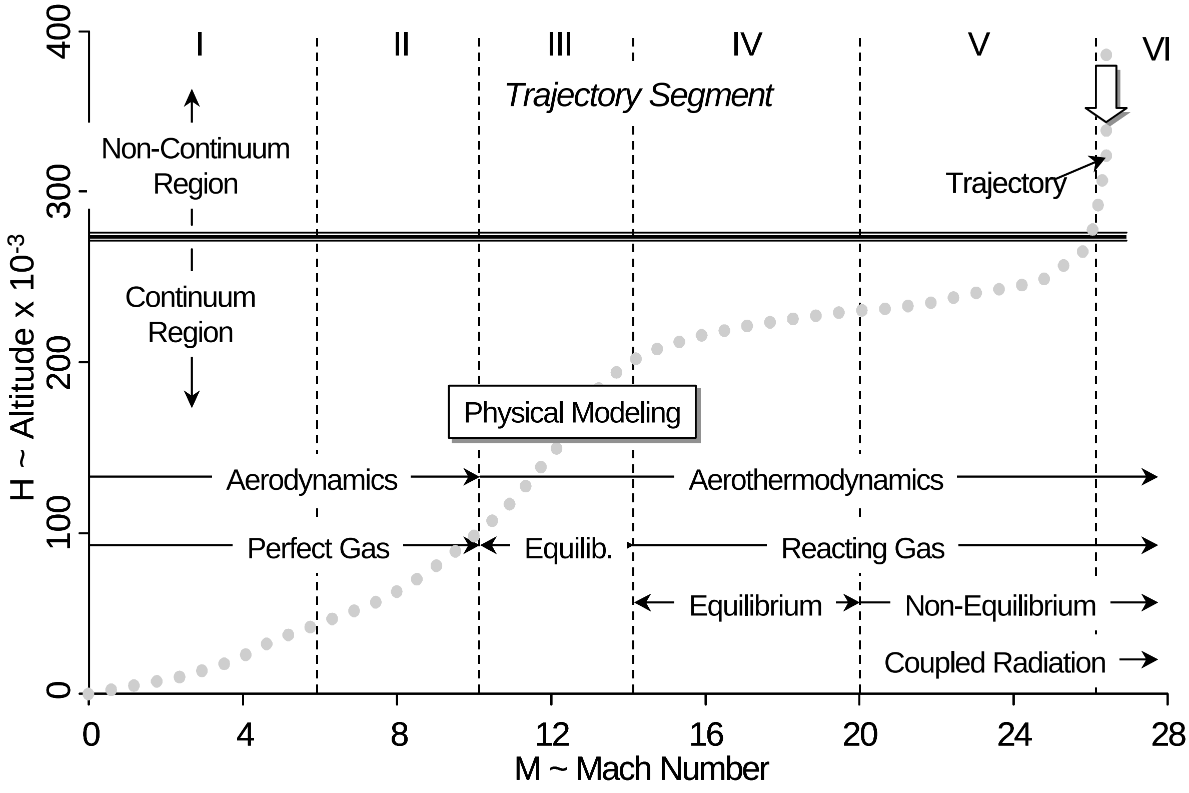
- ***Generation and management of large, complex data archives*** **that are shared across an entire community – e.g. DOE’s Human Genome Project data and NASA’s Earth Observing System data**

Grids Will Facilitate “Large-Scale” Problems

“Scale” has many dimensions.

- **Large scale parameter studies and data analysis frameworks: Thousands of processes tens of thousands of data files, petabyes of data**
  - **NASA’s reusable launch vehicle design studies**
  - **High Energy Physics data analysis**

Typical RLV Descent Trajectory: Aerodynamics Analyses

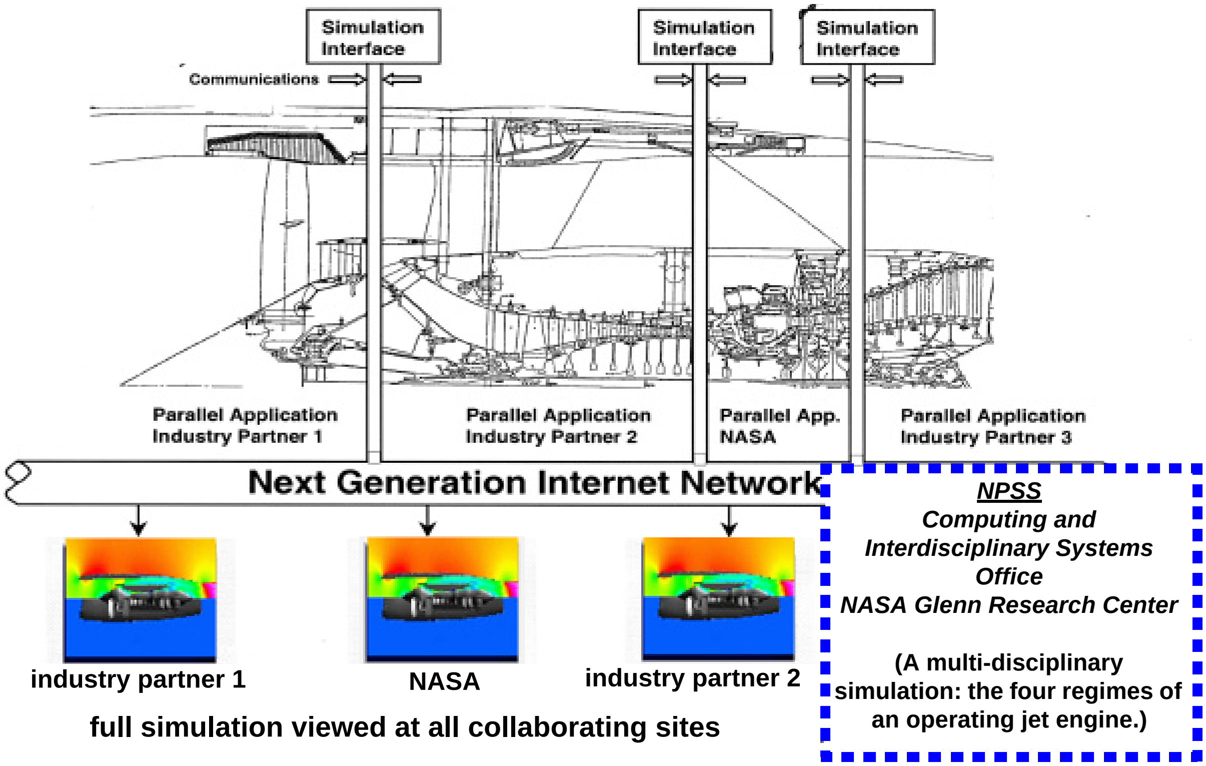


Scale in Grids (cont.)

- High computing capacity through the aggregation of many resource
  - coupled, multidisciplinary simulations too large for single systems
    - multiple model climate simulations
    - multi-component turbomachine simulation, e.g. in the Numerical Propulsion System Simulation (NPSS) program [3]

Collaborative Multi-Component Distributed Simulation

inlet compressor combustor turbine


Simulation Interface
Simulation Interface
Simulation Interface
Communications
Parallel Application Industry Partner 1
Parallel Application Industry Partner 2
Parallel App. NASA
Parallel Application Industry Partner 3
Next Generation Internet Network

industry partner 1 NASA industry partner 2

full simulation viewed at all collaborating sites

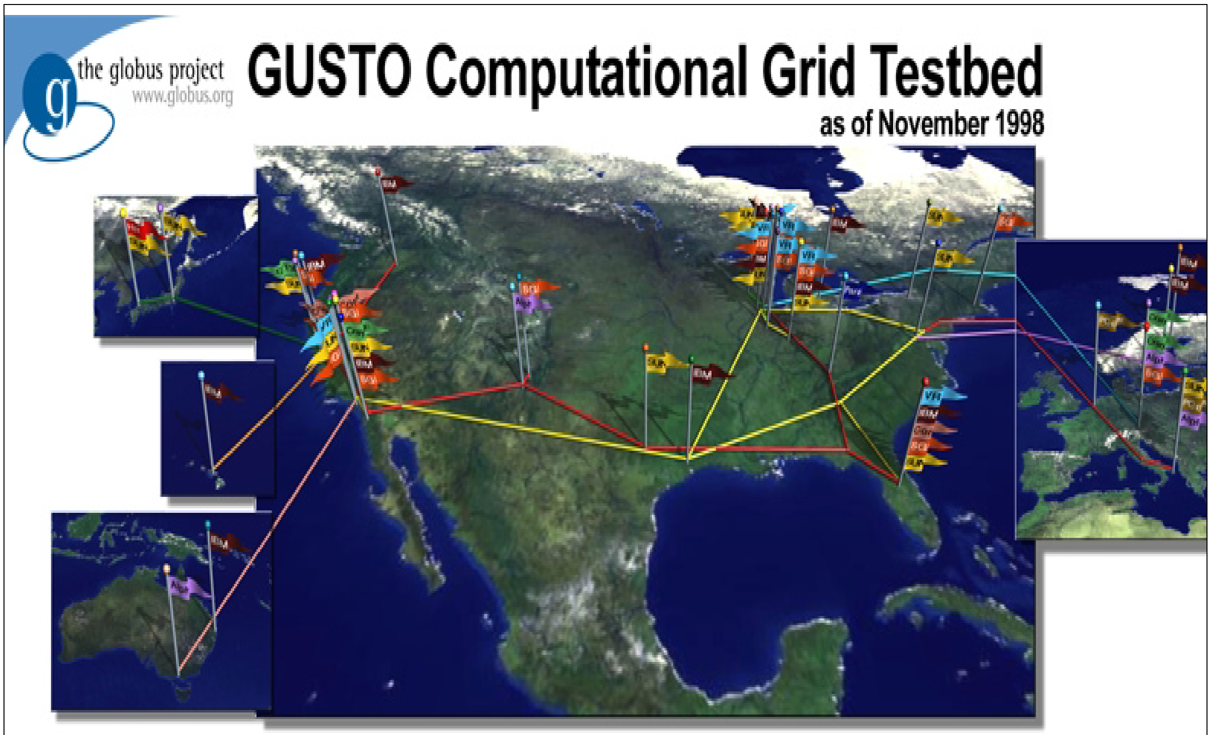
<u>NPSS</u>
Computing and Interdisciplinary Systems Office
NASA Glenn Research Center

(A multi-disciplinary simulation: the four regimes of an operating jet engine.)

Scale in Grids (cont.)

¨ **Heterogeneity and dispersed resources**

- **Grids are built from, and designed to manage, *organizationally and geographically dispersed components* (Grid management of multi-organization resources is designed to preserve local control of the resources)**
- **Grids are intended to provide *uniform views* of, and uniform access to many different resources for the computational scientist / code developer**
- **The GUSTO demonstration involved using 10s of supercomputers on four continents for a single problem**



Geographic and Organizational Complexity is the Norm in Grids.

Scale in Grids (cont.)

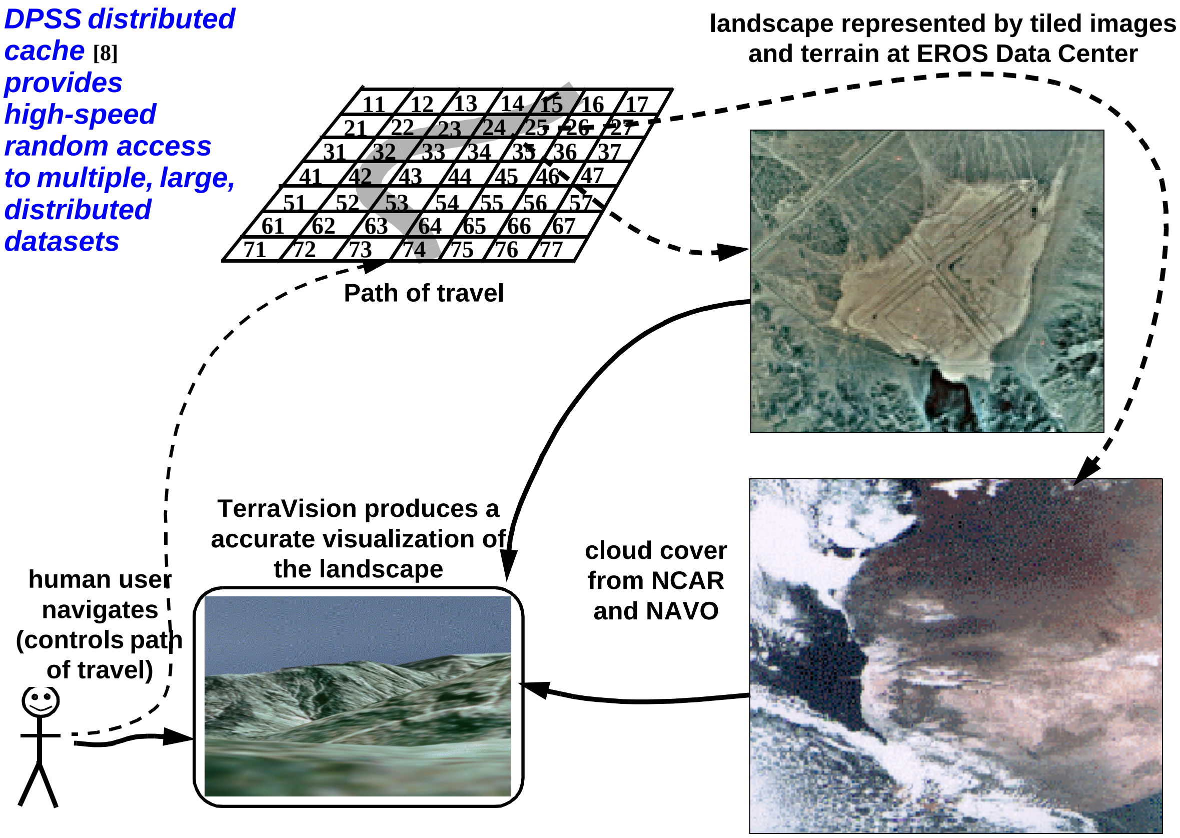
¨ **High data-rate, widely distributed data management**

- **MAGIC Gigabit Testbed: federated access for archived satellite and aerial imagery, digital terrain data, and atmospheric data [10], [11])**
  - **on demand, real-time interactive exploration of an operational environment supporting, e.g., military operations and community emergency services**
  - **aggregation of multiple, widely distributed, multi-discipline data sets**

MAGIC Testbed

- **on-line, real-time access to multiple environmental data sets that are (and always will be) maintained by domain experts at their own sites.**
- **DARPA MAGIC testbed consortium (see www.magic.net) developed distributed services, data and visualization from EROS Data Center, NCAR, NAVO, SRI**

- ***Similar characteristics to DOE's Combustion Corridor Project [21]***



DPSS distributed cache [8] ***provides high-speed random access to multiple, large, distributed datasets***

landscape represented by tiled images and terrain at EROS Data Center

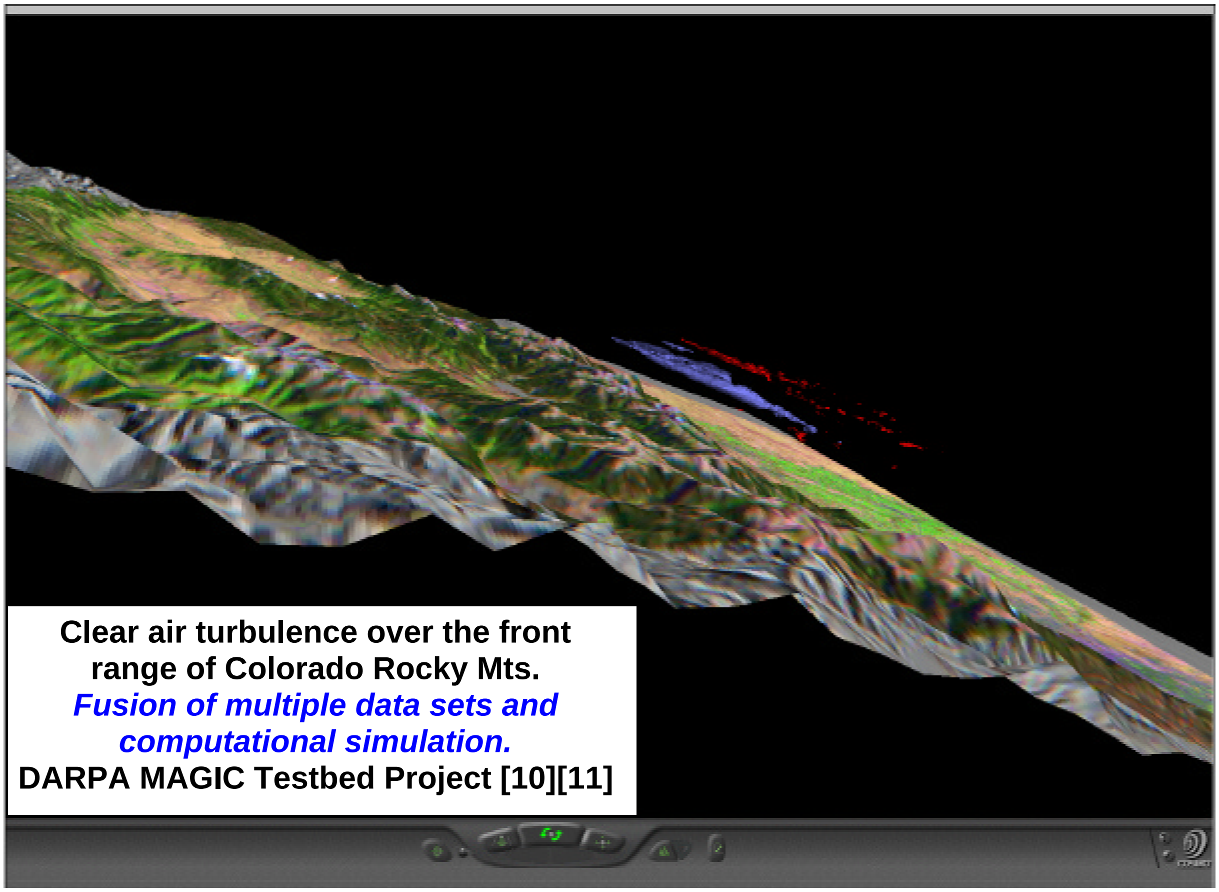
Path of travel

TerraVision produces a accurate visualization of the landscape

cloud cover from NCAR and NAVO

human user navigates (controls path of travel)

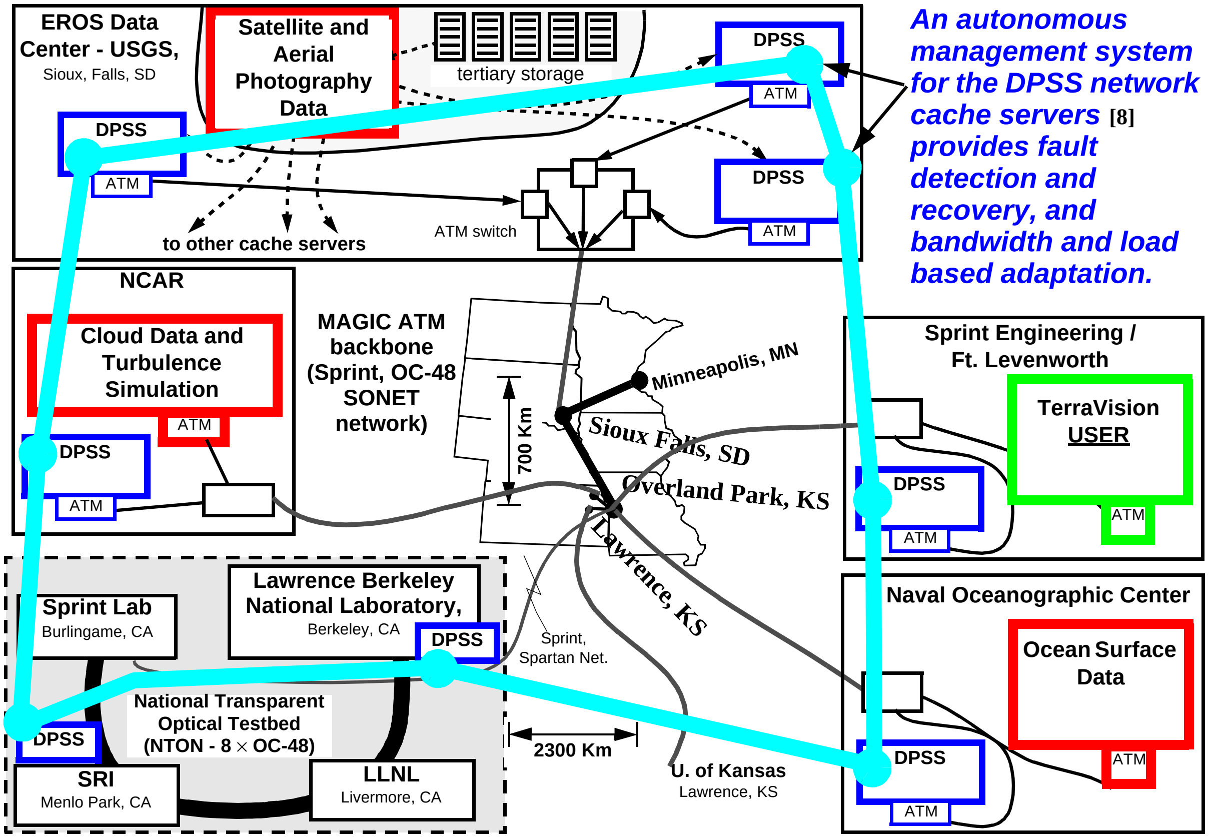
MAGIC and TerraVision Provide Real-time Visualization of Aggregated Data



Clear air turbulence over the front range of Colorado Rocky Mts.

Fusion of multiple data sets and computational simulation.

DARPA MAGIC Testbed Project [10][11]



An autonomous management system for the DPSS network cache servers [8] *provides fault detection and recovery, and bandwidth and load based adaptation.*

The MAGIC Testbed Distributed Application Environment

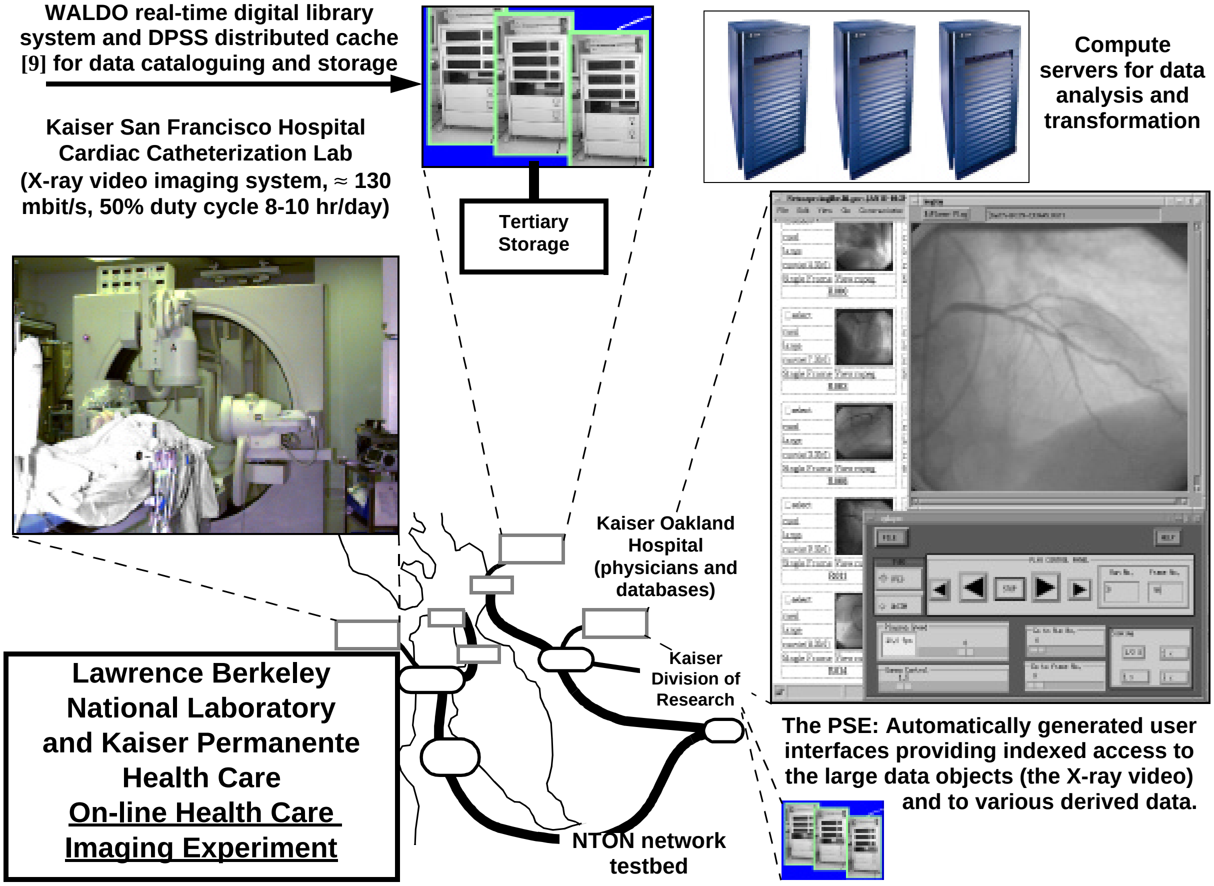
Scale in Grids (cont.)

¨ **Coupling large-scale computing and data systems to scientific and engineering instruments requires that *many heterogeneous resources be characterized, located, co-scheduled, and managed*, including collaborators, instruments, bandwidth, and storage and computational systems**

- **E.g.: real-time interaction with experiments through real-time data analysis and interpretation presented to the experimentalist in ways that allow direct interaction with the experiment (instead of just with the instrument controls)**
- **E.g.: real-time processing and distribution of satellite data feeds**

Scale in Grids (cont.)

- **On-line medical imaging system (*real-time digital libraries for on-line, high data-rate instruments* [9])**
  - **on-line, real-time, high data-rate medical instrument with remote users**
  - **distributed data analysis and automatic data cataloguing and archiving**
  - **strict authorization and access control**
  - **optical WDM metropolitan area network (NTON)**



WALDO real-time digital library system and DPSS distributed cache [9] for data cataloguing and storage

Kaiser San Francisco Hospital Cardiac Catheterization Lab (X-ray video imaging system, ≈ 130 mbit/s, 50% duty cycle 8-10 hr/day)

Tertiary Storage

Compute servers for data analysis and transformation

Kaiser Oakland Hospital (physicians and databases)

Kaiser Division of Research

NTON network testbed

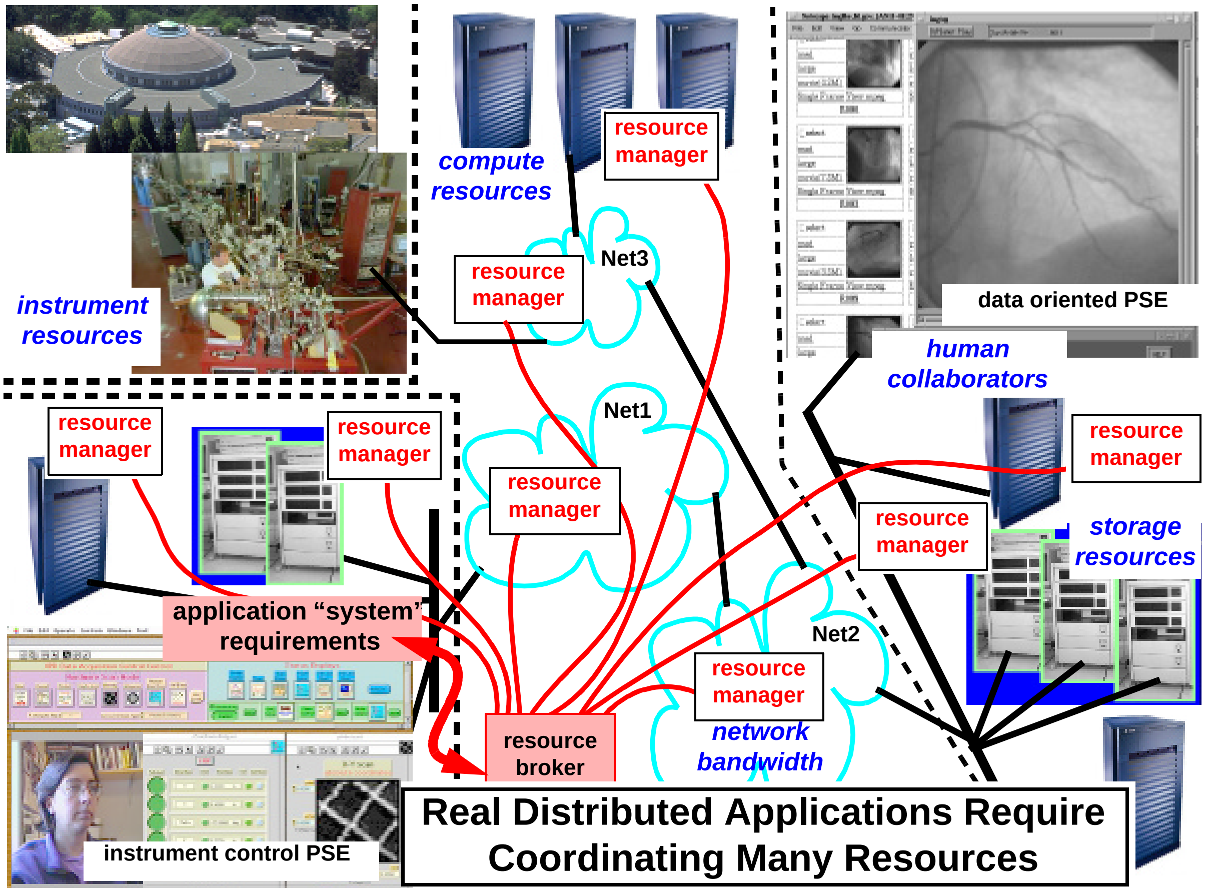
Lawrence Berkeley National Laboratory and Kaiser Permanente Health Care
<u>On-line Health Care Imaging Experiment</u>

The PSE: Automatically generated user interfaces providing indexed access to the large data objects (the X-ray video) and to various derived data.

This integration of scientific and engineering instruments with large-scale computing and data systems involves many different types of resources that are accessed and shared across many institutions. This implies:

- ***numerous interconnected servers providing computational simulation and analysis, data access, and functional access to instruments, in research networks (e.g., ESNet, NREN, Internet-2)***
- ***many simultaneous collaborators, e.g. at DOE Labs, NASA Centers, other Federal labs, industrial partners, and many universities***

and resource management becomes a dominate problem.



Real Distributed Applications Require Coordinating Many Resources

Scale in Grids (cont.)

¨ **Single computational problems too large for any single system**

- **E.g.: An aerodynamics reference calculation – sufficiently fine geometry that solution no longer changes with cell size – that requires terabytes of memory.**

Requirements

Analysis of these types of science and engineering scenarios provide various requirements for Grids.

¨ **Generation and management of large, complex data archives – e.g. geometry (structure) and performance of airframes and turbomachines – that are maintained by discipline experts at different sites, and must be *accessed and updated by collaborating analysts*.**

¨ ***Complex workflow scenarios* involving many compute and data intensive steps must be managed – e.g. in mutli-disciplinary simulations and data analysis protocols.**

- Existing heterogeneous *simulation components need to be coupled and operated simultaneously* in order to provide whole system simulations (e.g. “multi-disciplinary optimization”).
- Interfaces to computational and data analysis tools must provide *appropriate levels of abstraction* for discipline problem solvers.
- Techniques are needed to *search, interpret, and fuse data from multiple remote archives*.
- Scientists and engineers must be able to *securely and selectively share* all aspects of their work process.

Requirements (cont.)

- *Data streams from instrument systems* must be available in real-time for computational data analysis systems, and for cataloguing and entry into databases.
- In widely distributed environments, tools and services for autonomous *fault management* and recovery are required for both applications and infrastructure
- *Uniform access* and management for many heterogeneous resources

Requirements (cont.)

These general requirements imply certain capabilities that must be provided by Grids in order to support the interactions of the instruments, people, information systems, and computing resources needed to facilitate distributed science and engineering.

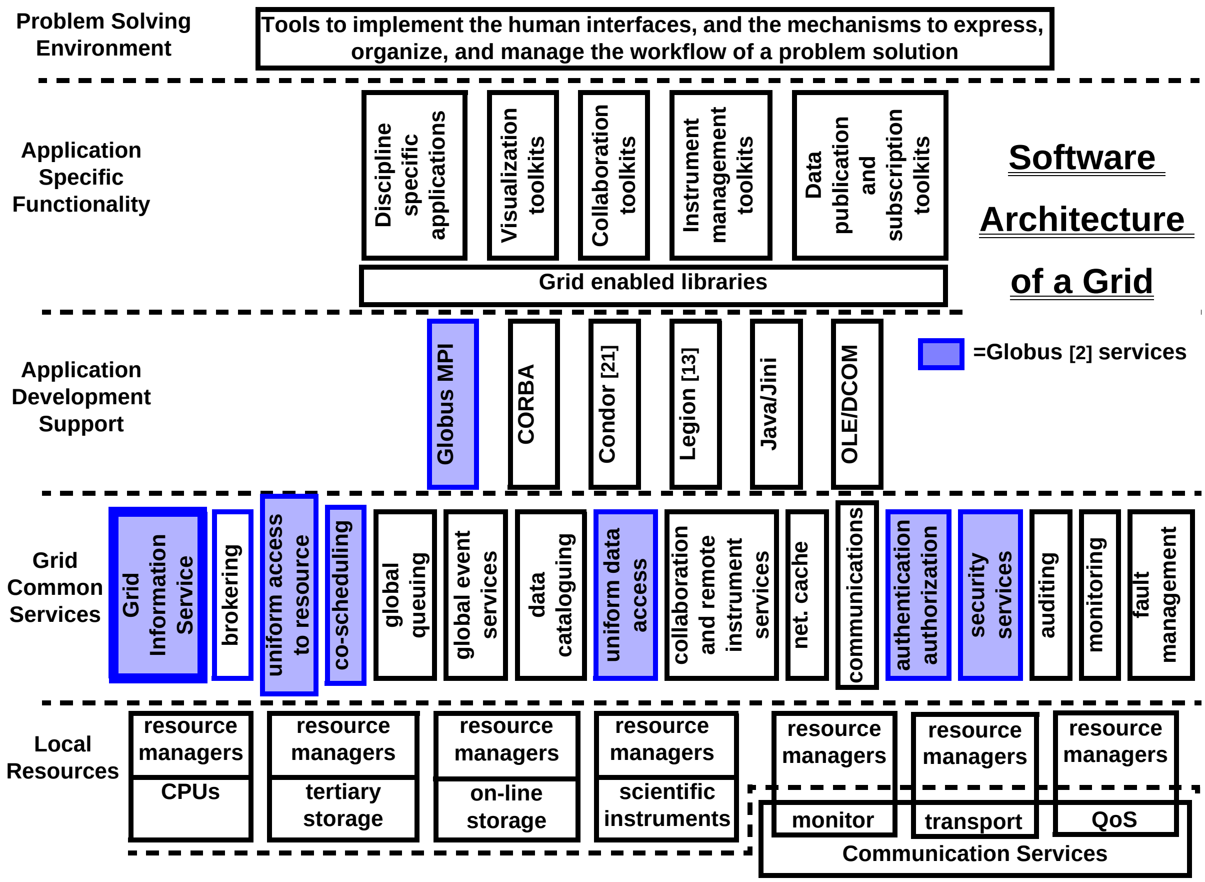
There is an additional set of requirements from the tool developers – the application domain computational scientists – primarily in the area of supporting code development and execution environment access and management. These will not be discussed here, but are also major drivers for developing Grid functionality.

What are "Grids"?

Science / Computation, Data, and Collaboration Grids are distributed, high performance computing and data handling infrastructure that

- is persistent and supported
- incorporates geographically and organizationally dispersed, heterogeneous resources, including
  - computing systems
  - storage systems
  - instruments and other real-time data sources
  - human collaborators
  - communications systems
- provides common interfaces for all resources
- support resource aggregation

Software Architecture of a Grid



Problem Solving Environment: Tools to implement the human interfaces, and the mechanisms to express, organize, and manage the workflow of a problem solution

Application Specific Functionality: Discipline specific applications; Visualization toolkits; Collaboration toolkits; Instrument management toolkits; Data publication and subscription toolkits; Grid enabled libraries

Application Development Support: Globus MPI; CORBA; Condor [21]; Legion [13]; Java/Jini; OLE/DCOM

🟦 =Globus [2] services

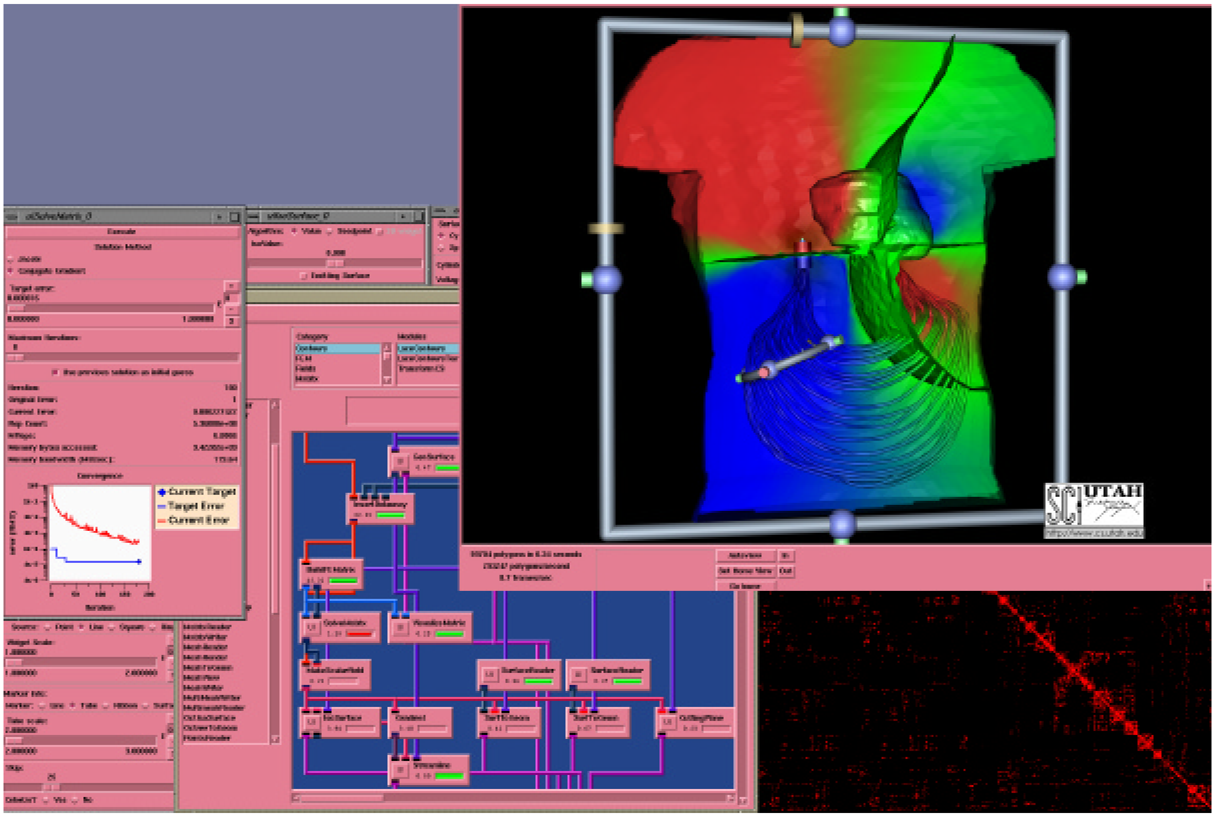
Grid Common Services: Grid Information Service; brokering; uniform access to resource; co-scheduling; global queuing; global event services; data cataloguing; uniform data access; collaboration and remote instrument services; net. cache; communications; authentication authorization; security services; auditing; monitoring; fault management

Local Resources: resource managers / CPUs; resource managers / tertiary storage; resource managers / on-line storage; resource managers / scientific instruments; resource managers / monitor; resource managers / transport; resource managers / QoS; Communication Services

What Must Grids Provide?

To satisfy requirements arising from scientific and engineering applications, certain functionality must be provided. *Example* functionality and what it facilitates includes:

- **Modular toolkits for building Problem Solving Environments/Frameworks that provide workflow management, application code composition, access control, and collaboration**
  - **the “user interface to the Grid”**
  - **knowledge based workflow control**
  - **high throughput job managers (specialized PSEs for parameter studies)**



Example Problem Solving Environment. [24]

What Must Grids Provide (cont.)

¨ **Support for multiple programming environments**

Þ **provide various approaches for building applications that use distributed computing systems, federated data sources, and other distributed resources**

¨ **Resource discovery and brokering, advance reservation, and co-scheduling for all resources**

Þ **large-scale computing through aggregation**

Þ **on-demand and scheduled, dynamic system construction**

Þ **facilitates fault recovery**

¨ **Monitoring for performance tuning,
fault detection and recovery, and management**

Þ **construction of reliable, production quality applications**

Þ **supports autonomous system management**

What Must Grids Provide (cont.)

¨ **End-to-end high bandwidth between resources**

Þ **support for high data-rate distributed applications**

Þ **coupling of remote instruments with large-scale computing**

Þ **accommodation of very long round-trip-time communication**

¨ **Use of multi-source data resources**

Þ **federating datasets to support multi-disciplinary simulation**

¨ **Data location management and optimized remote data access**

Þ **automatic location management to minimize data movement and data transfer time when CPUs and data archives are in different geographic locations**

What Must Grids Provide (cont.)

¨ **Global event management**

Þ **synchronization of widely distributed processes and data sets to support consistency/repeatability of results, and use of “independent” data sources**

Þ **primary information source for workflow management**

¨ **Grid enabled / aware algorithms**

Þ **distributing single codes across Grids requires new techniques**

¨ **Security and infrastructure protection**

Þ **assurance of resource use and function in the semi-open environment of science and engineering R&D environments**

Þ **secure autonomous operation**

What Must Grids Provide (cont.)

- ¨ **Access control mechanisms**
  - Þ **management of access rights by data / resource stakeholders**
  - Þ **positive identification of users and management of user attributes pertaining to access rights**
- ¨ **Operational procedures and tools**
  - Þ **management of cross-organizational resources**
  - Þ **management of widely distributed resources**
- ¨ **User support, allocation management, and accounting**

Vision for Science Grids

Computing and data Grids in the service of science will provide significant new capabilities to scientists and engineers by facilitating *routine construction of information based problem solving environments / frameworks that knit together widely distributed computing, data, and instrument systems*
– esp. supercomputers, petabyte storage systems, and unique national-scale instruments –
together with human resources, *into aggregated systems that can address complex and large-scale computing and data analysis problems* beyond what is possible today.

Grids also have the potential to *provide pools of resources that could be called on in extraordinary / rapid response situations* (such as disaster response) because they can provide:

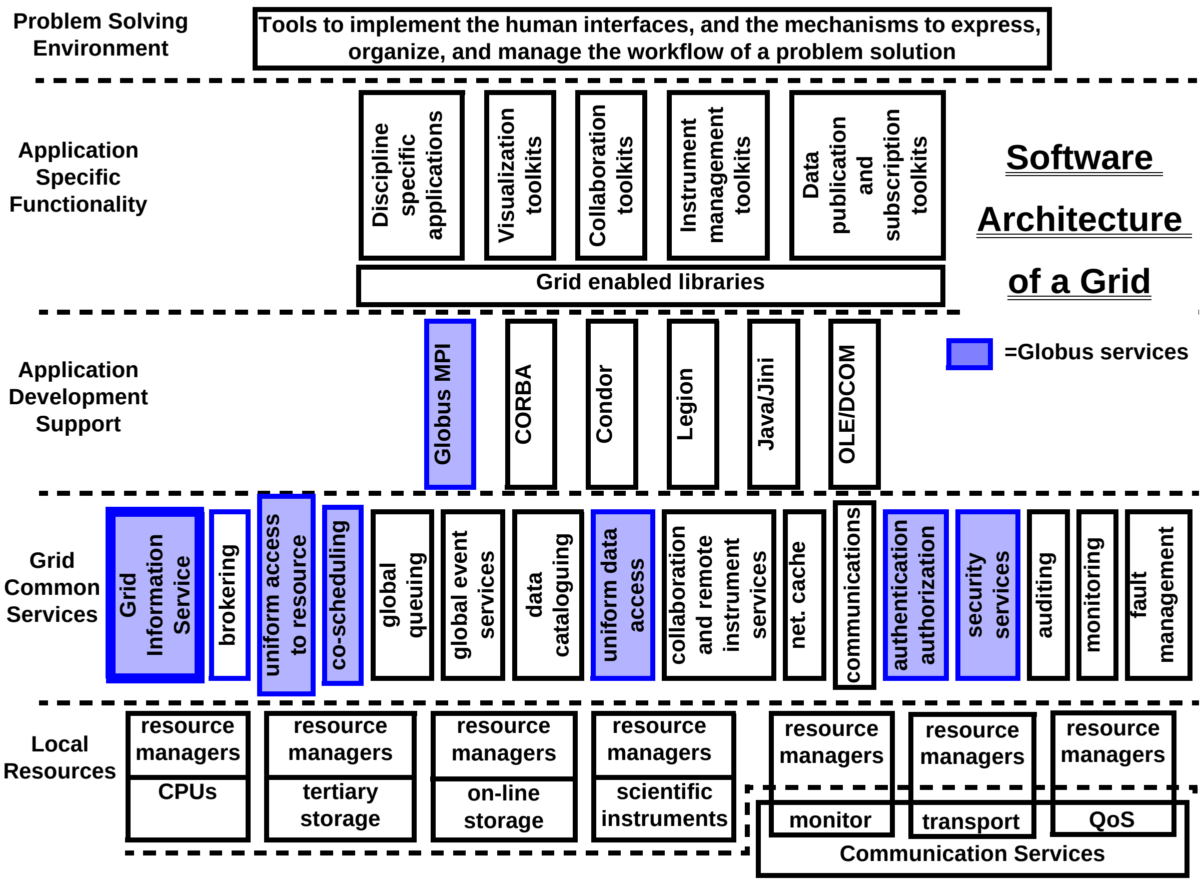
- common interfaces and access mechanisms
- standardized management
- uniform user authentication and authorization

for large collections of distributed resources (whether or not they normally function in concert).

The Implementation of a Grid

- *A conceptual framework for describing / organizing* the functional components (see figure)
- Toolkits for building *PSEs / Frameworks*
- Multiple *middleware* systems (code development)
- Grid enabled *toolkits / libraries*
- *Grid Common Services* (mostly resource access and management related) – mostly Globus [2]
- *Resources* (compute, data, instruments, humans)
- *Operational infrastructure* (e.g., auditing, security, access control, user and system support)
- *Testbeds and prototypes*
- *R&D* for new capabilities

Software Architecture of a Grid



Problem Solving Environment

Tools to implement the human interfaces, and the mechanisms to express, organize, and manage the workflow of a problem solution

Application Specific Functionality

Discipline specific applications

Visualization toolkits

Collaboration toolkits

Instrument management toolkits

Data publication and subscription toolkits

Grid enabled libraries

Application Development Support

Globus MPI

CORBA

Condor

Legion

Java/Jini

OLE/DCOM

🟦 =Globus services

Grid Common Services

Grid Information Service

brokering

uniform access to resource

co-scheduling

global queuing

global event services

data cataloguing

uniform data access

collaboration and remote instrument services

net. cache

communications

authentication authorization

security services

auditing

monitoring

fault management

Local Resources

resource managers / CPUs

resource managers / tertiary storage

resource managers / on-line storage

resource managers / scientific instruments

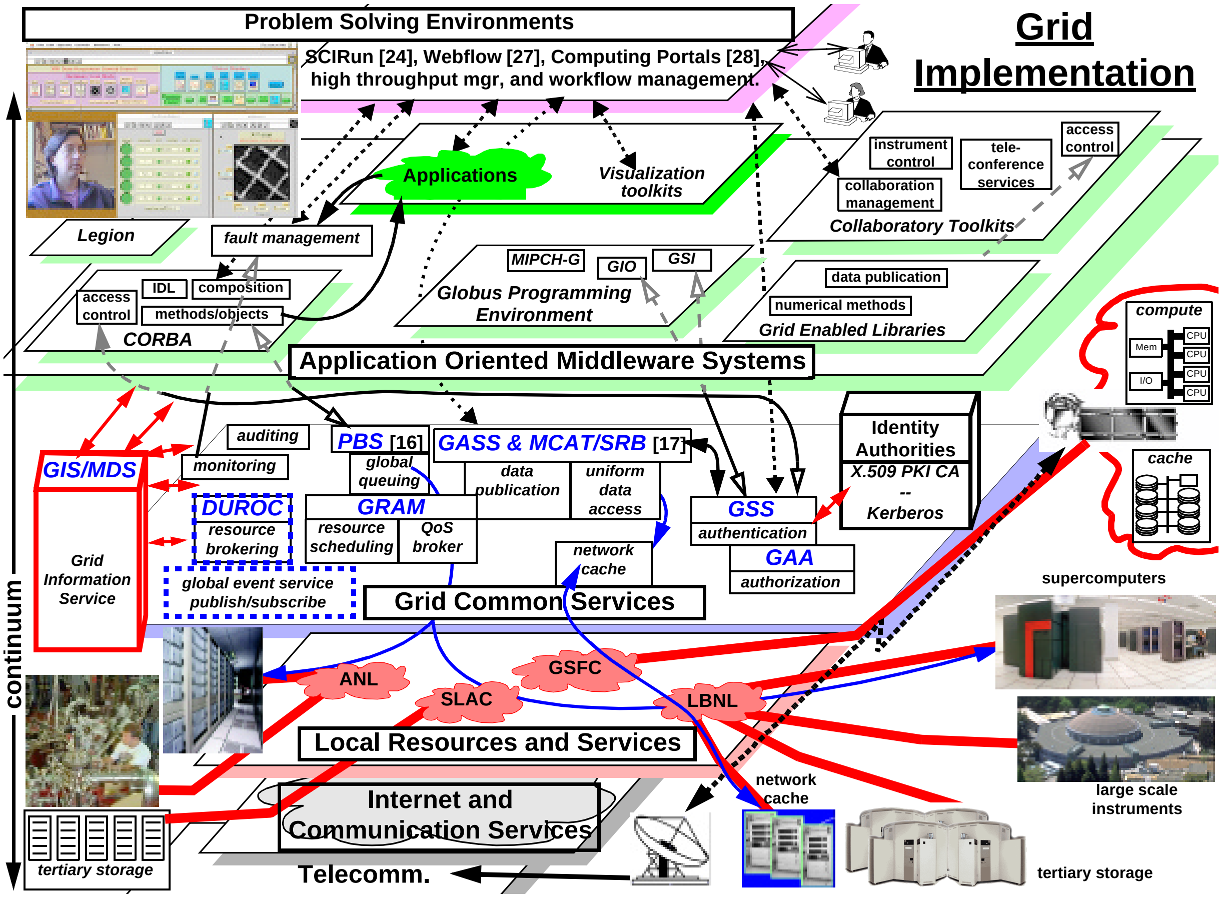
resource managers / monitor

resource managers / transport

resource managers / QoS

Communication Services

Grid Implementation



Problem Solving Environments

SCIRun [24], Webflow [27], Computing Portals [28], high throughput mgr, and workflow management.

Applications

Visualization toolkits

instrument control

tele-conference services

access control

collaboration management

Collaboratory Toolkits

Legion

fault management

access control

IDL

composition

methods/objects

CORBA

MIPCH-G

GIO

GSI

Globus Programming Environment

data publication

numerical methods

Grid Enabled Libraries

Application Oriented Middleware Systems

auditing

monitoring

GIS/MDS

Grid Information Service

DUROC

resource brokering

global event service publish/subscribe

PBS [16]

global queuing

GRAM

resource scheduling

QoS broker

GASS & MCAT/SRB [17]

data publication

uniform data access

network cache

Grid Common Services

GSS

authentication

GAA

authorization

Identity Authorities

X.509 PKI CA

--

Kerberos

compute

Mem

I/O

CPU

CPU

CPU

CPU

cache

supercomputers

ANL

GSFC

SLAC

LBNL

Local Resources and Services

large scale instruments

network cache

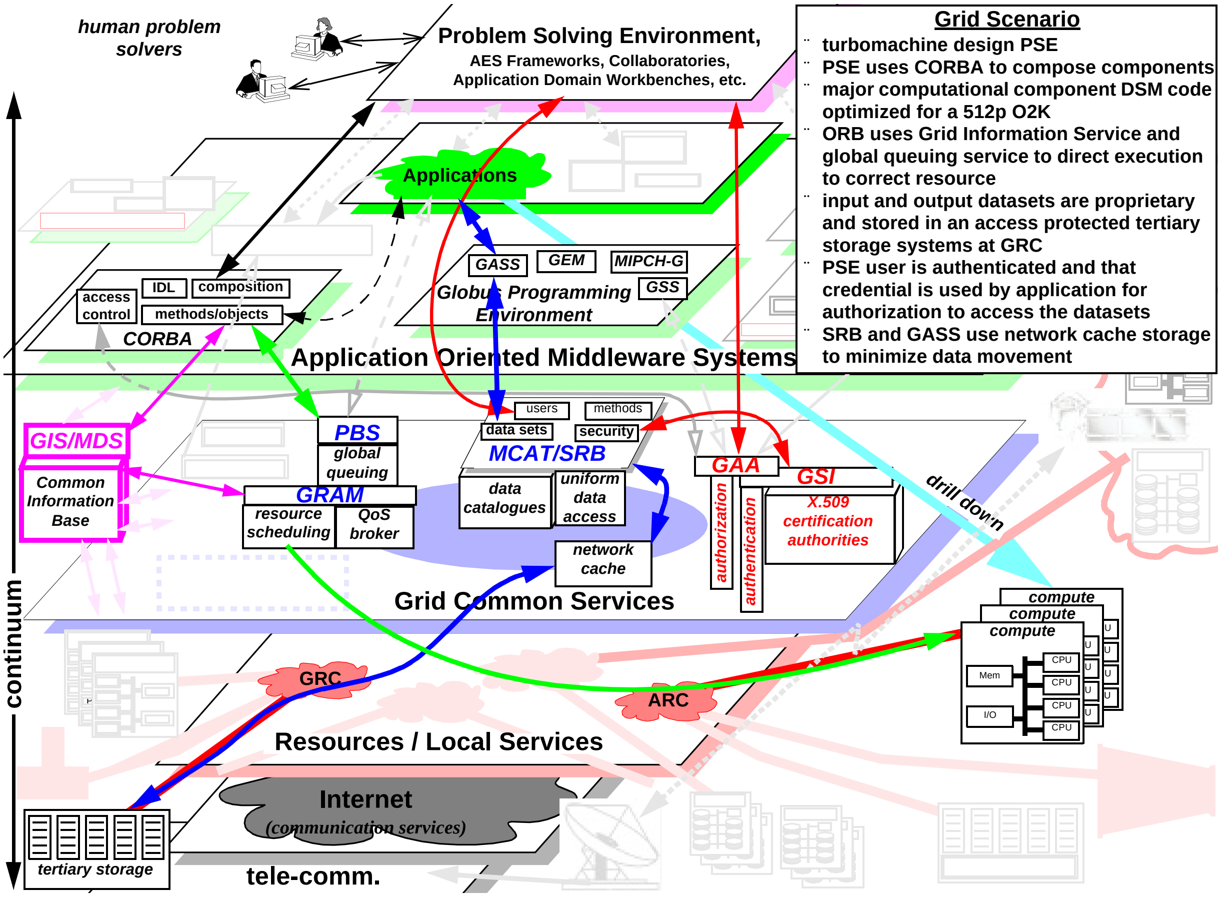
Internet and Communication Services

Telecomm.

tertiary storage

tertiary storage

continuum



Grid Scenario

- turbomachine design PSE
- PSE uses CORBA to compose components
- major computational component DSM code optimized for a 512p O2K
- ORB uses Grid Information Service and global queuing service to direct execution to correct resource
- input and output datasets are proprietary and stored in an access protected tertiary storage systems at GRC
- PSE user is authenticated and that credential is used by application for authorization to access the datasets
- SRB and GASS use network cache storage to minimize data movement

Information Power Grid

Distributed High-Performance Computing and Large-Scale Data Management for Science and Engineering

William E. Johnston,

Dennis Gannon, and Bill Nitzberg

Numerical Aerospace Simulation Division

William J. Feiereisen, Division Chief,

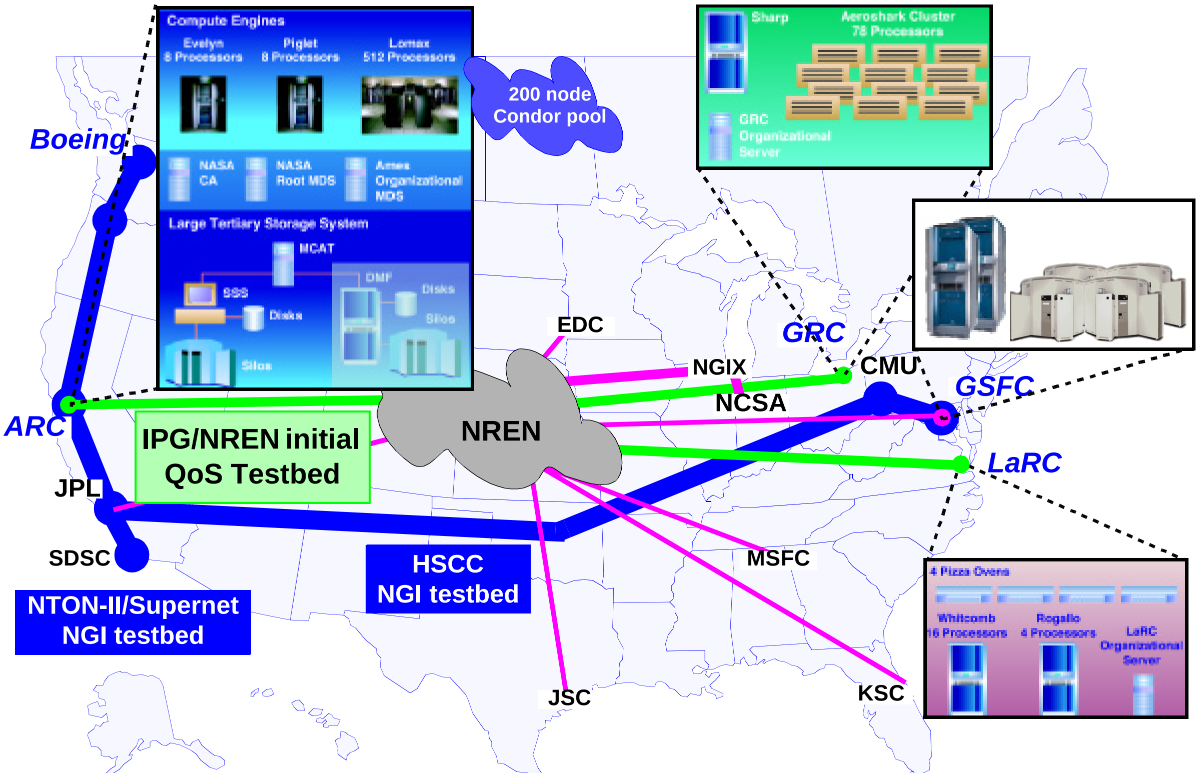
William Thigpen, Engineering Branch Chief,

Alex C. Woo Research Branch Chief

http://www.nas.nasa.gov/IPG

How is IPG Being Accomplished?

- NASA: IT/ACNS, HPCC/CAS, and HPCC/NREN
- Collaboration between
  - several NASA centers (Ames/NAS, GRC, LaRC, GSFC and JPL)
  - the NSF PACIs (NCSA / Alliance and SDSC / NPACI)
  - universities and government labs
- Areas that must be addressed for baseline IPG:
  1) persistent operational environment that encompasses significant resources
  2) new service delivery / operational models
  3) new functionality



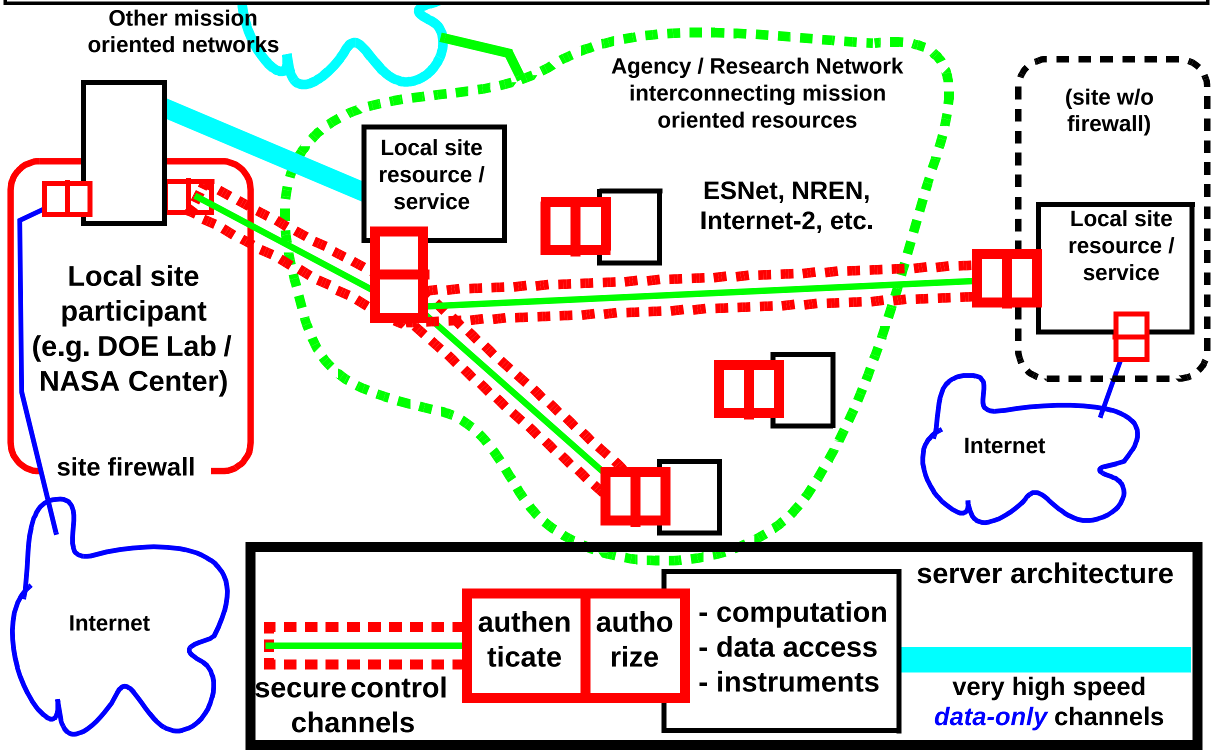
IPG Baseline System and
High Data-Rate (DX) Testbed

Access Control and Security

A security model for the Grid must address:

- cyber risk mitigation and cross-site integrity
- control channel integrity and confidentiality
- optional data channel integrity and confidentiality
- identity management, authentication, and single identity sign-on w/o clear text passwords [20]
- authorization via policy-based access control
- infrastructure assurance

Security Model: All Grid services command and control functions are transported over an encrypted channel, after the client/user is authenticated and authorized. This compartmentalizes all servers: If multiple servers are involved in a distributed system, then each reauthorizes connections through the use of cryptographic proxies or active re-authentication.

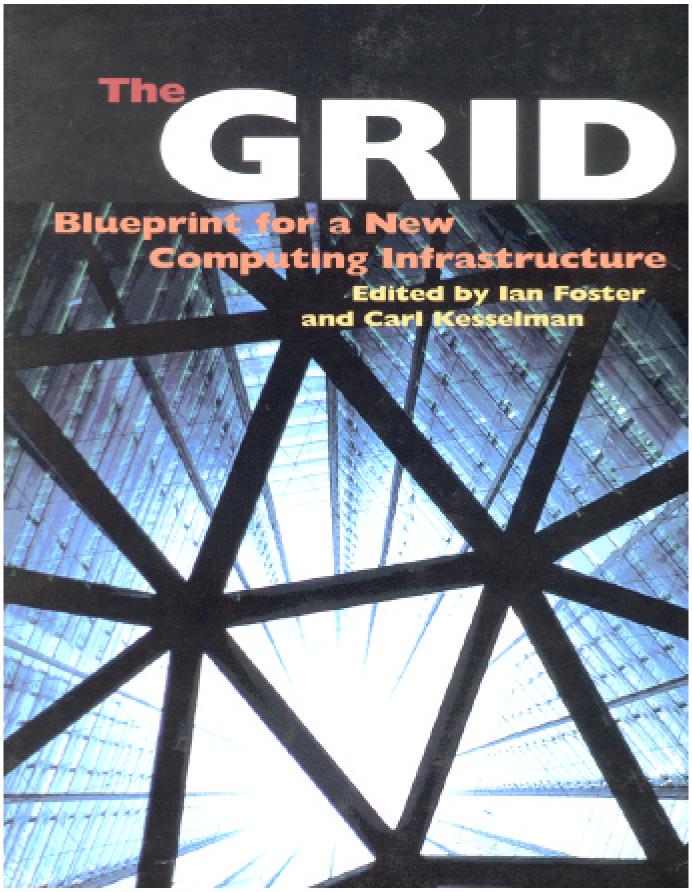

Other mission oriented networks
Agency / Research Network interconnecting mission oriented resources
(site w/o firewall)
Local site resource / service
ESNet, NREN, Internet-2, etc.
Local site resource / service
Local site participant (e.g. DOE Lab / NASA Center)
site firewall
Internet
Internet
server architecture
authen ticate
autho rize
- computation
- data access
- instruments
secure control channels
very high speed *data-only* channels

The Components of a Grid (cont.)

Internet / Telecommunications (e.g., DOE environment)

- **ESNet [29] network infrastructure will support the DOE Science Grid distributed computing testbed and bandwidth reservation R&D.**
- **ESNet high-speed R&D links and DARPA NGI “Supernet” testbed (mixed optical WDM @ 10 Gb/s and SONET OC-48 POS - 2.5 Gb/s) will provide the DOE Science Grid high data-rate testbed.**

The same situation is true for NASA’s NREN [30] and IPG.



This book [3] examines some of the technologies and issues for computational and data Grids.

For ongoing work in this area see www.gridforum.org [14]

References and Notes

[1] The Institute of Electrical and Electronic Engineers International Symposium on High Performance Distributed Computing (HPDC) provides a forum for presenting the latest research findings that unify parallel and distributed computing. In HPDC environments, parallel or distributed computing techniques are applied to the solution of computationally intensive applications across networks of computers.

[2] Globus is a middleware system that provides a suite of services designed to support high performance, distributed applications. Globus provides:

- Resource Management: Components that provide standardized interfaces to various local resource management systems (GRAM) manage allocation of collections of resources (DUROC). All Globus resource management tools are tied together by a uniform resource specification language (RSL).
- Remote Access: Components that enable remote access to files (GASS and RIO) and executables (GEM).
- Security: Support for single sign-on, authentication, and authorization within the Globus system (GSI) and (experimentally) authorization (GAA).
- Fault Detection: Basic support for building fault detection and recovery into Globus applications.
- Information Infrastructure: Global access to information about the state and configuration of system components of an application (MDS).

- Grid programming services: Support writing parallel-distributed programs (MPICH-G), monitoring (HBM), etc.

www.globus.org provides full information about the Globus system.

[3] *The Grid: Blueprint for a New Computing Infrastructure*, edited by Ian Foster and Carl Kesselman. Morgan Kaufmann, Pub. August 1998. ISBN 1-55860-475-8. http://www.mkp.com/books_catalog/1-55860-475-8.asp

[4] “Grids as Production Computing Environments: The Engineering Aspects of NASA's Information Power Grid,” William E. Johnston, Dennis Gannon, and Bill Nitzberg. Eighth IEEE International Symposium on High Performance Distributed Computing, Aug. 3-6, 1999, Redondo Beach, California. (Available at http://www.nas.nasa.gov/~wej/IPG)

[5] “The Diesel Combustion Collaboratory will facilitate collaboration among the participants in a distributed research project, specifically the Heavy Duty Diesel Combustion CRADA (Cooperative Research and Development Agreement).” See http://www-collab.ca.sandia.gov

[6] See www.nas.nasa.gov/IPG for project information and pointers.

[7] “Numerical Propulsion System Simulation (NPSS) is a concerted effort by NASA Glenn Research Center, the aerospace industry and academia to develop an advanced engineering environment - or integrated collection of software programs - for the analysis and design of aircraft engines and, eventually, space transportation components. Its purpose is to dramatically reduce the time, effort and expense necessary to design and test jet engines. It accomplishes that by generating sophisticated computer simulations of an aerospace object or system, thus permitting an engineer to “test” various design options without having to conduct costly and time-consuming real-life tests. The ultimate goal of NPSS is to create a “numerical test cell” that enables engineers to create complete engine simulations overnight on cost-effective computing platforms.” See http://hpcc.grc.nasa.gov/hpcc2/npssintro.shtml

[8] Tierney, B. Lee, J., Crowley, B., Holding, M., Hylton, J., Drake, F., “A Network-Aware Distributed Storage Cache for Data Intensive Environments”, Proceeding of IEEE High Performance Distributed Computing conference (HPDC-8), August 1999.

[9] “Real-Time Generation and Cataloguing of Large Data-Objects in Widely Distributed Environments,” W.Johnston, Jin G., C. Larsen, J. Lee, G. Hoo, M. Thompson, and B. Tierney (LBNL) and J. Terdiman (Kaiser Permanente Division of Research). Invited paper, International Journal of Digital Libraries - Special Issue on “Digital Libraries in Medicine”. May, 1998. http://www-itg.lbl.gov/WALDO/

[10] MAGIC: “The MAGIC Gigabit Network.” See: http://www.magic.net

[11] TerraVision-2: VRML based data fusion and browsing. (MAGIC consortium, NCAR, and NAVO: http://www.ai.sri.com/TerraVision/)

[12] Clipper: The goal of the Clipper project is software systems and testbed environments that result in a collection of independent but architecturally consistent service components that will enhance the ability of applications and systems to construct and use widely distributed, high-performance data and computing infrastructure. Such middleware should support high-speed access and integrated views for multiple data archives; resource discovery and automated brokering; comprehensive real-time monitoring and performance trend analysis of the networked subsystems, including the storage, computing, and middleware components, and; flexible and distributed management of access control and policy enforcement for multi-administrative domain resources. See http://www-itg.lbl.gov/~johnston/Clipper

[13] Legion is an object-based, meta-systems software project at the University of Virginia. http://www.cs.virginia.edu/~legion/

[14] The Grid Forum (www.gridforum.org) is an informal consortium of institutions and individuals working on wide area computing and computational grids: the technologies that underlay such activities as the NCSA Alliance's National Technology Grid, NPACI's Metasystems efforts, NASA's Information Power Grid, DOE ASCI's DISCOM program, and other activities worldwide.

[15] “The ISE Goal: To develop the capability for scientists and engineers to work together in a virtual environment, using simulation to model the complete life-cycle of a product/mission before commitments are made to produce physical products.” www.ise.nasa.gov

[16] The Portable Batch System (PBS) is a batch queueing system developed at NAS. PBS implements the POSIX standard, and operates on networked, multi-platform UNIX environments, including heterogeneous clusters of workstations, supercomputers, and massively parallel systems. PBS is the basis of the IPG global queuing system work. http://parallel.nas.nasa.gov/Parallel/PBS/

[17] Storage Resource Broker (SRB) provides uniform access mechanism to diverse and distributed data sources. The SRB provides the protocol conversion required to interface to heterogeneous data sources. The SRB interfaces with the MCAT metadata catalog to access metadata information about individual files and objects. http://www.sdsc.edu/MDAS/

[18] Alliance / NCSA: National Computational Science Alliance. “The National Computational Science Alliance (Alliance) is [an NSF funded] partnership among computational scientists, computer scientists, and professionals in education, outreach, and training at more than 50 U.S. universities and research institutions working to prototype the computational and information infrastructure of the next century.” www.ncsa.edu

[19] NPACI: The National Partnership for Advanced Computational Infrastructure. “Led by UC San Diego, NPACI [an NSF funded partnership] will revolutionize the nation’s computational infrastructure by building on the foundation of SDSC and involving 37 of the nation's leading academic and research institutions, located in 18 states from coast to coast. Driven by real applications needs, NPACI is

developing software infrastructure to link the highest performance computers, data servers, and archival storage systems to enable easier use of the aggregate computing power.” www.npaci.edu

[20] PKI: Public-Key certificate Infrastructure. Public-key cryptography involves two keys, whereby data encrypted with one key can only be decrypted with the other, and visa versa. In PKI one key (the public-key) is freely available and the other is kept private. In this way, material encrypted with the private key and decrypted with the public-key proves that the holder of the private key must have been the originator of the material. A certification authority generates a certificate containing the name (usually X.500 distinguished name) of an entity (e.g. user) and that entity’s public key. The CA then signs this “certificate” and publishes it (usually in an LDAP directory service). These are the basic components of PKI, and allow the entity to prove its identity, independent of location or system, by signing a token with the private key, handing the signed token to a system (e.g. as part of a login process), and then that system can verify the signer’s identity by obtaining the identity certificate, extracting the entity’s public key, and verifying the signature. The identity certificate (most commonly an X.509 certificate) is, in turn, verified by obtaining the CA’s public key and using it to verify the contents.This later process is called digital signature and is accomplished by the certificate originator generating a unique hash of the certificate contents, and then encrypting that hash with the originator’s private key. The hash is then appended to the certificate (or any other document) and may be used to both verify the originator’s identity and the integrity of the contents (the hash function produces a “unique” hash for every byte string). For more information, see, e.g.,RSA Lab's "Frequently Asked Questions About Today's Cryptography" http://www.rsa.com/rsalabs/faq/, *Computer Communications Security: Principles, Standards, Protocols, and Techniques*. W. Ford, Prentice-Hall, Englewood Cliffs, New Jersey, 07632, 1995, or *Applied Cryptography*, B. Schneier, John Wiley & Sons, 1996.

[21] Condor is a High Throughput Computing environment that can manage very large collections of distributively owned workstations. The environment is based on a novel layered architecture that enables it to provide a powerful and flexible suite of Resource Management services to sequential and parallel applications. http://www.cs.wisc.edu/condor/

[22] NREN: NASA Research and Education Network. NREN is an integral partner in IPG, and supplies the IPG network testbeds and does network bandwidth reservation R&D. http://www.nren.nasa.gov

[23] RLV: Reusable Launch Vehicle. NASA's next generation space shuttle design program.

[24] SCIRun is a scientific programming environment that allows the interactive construction, debugging and steering of large-scale scientific computations. SCIRun can be used for interactively:

- Changing 2D and 3D geometry models (meshes).
- Controlling and changing numerical simulation methods and parameters.
- Performing scalar and vector field visualization.

SCIRun uses a visual programming dataflow system. SCIRun is extensible to a

variety of applications and will work with third party modules written in Fortran, C, and C++. http://www.cs.utah.edu/~sci/software/

[25] "The Network Weather Service is a distributed system that periodically monitors and dynamically forecasts the performance various network and computational resources can deliver over a given time interval." See http://nws.npaci.edu/

[26] "The NetLogger Methodology for High Performance Distributed Systems Performance Analysis," Brian Tierney, W. Johnston, J. Lee, G. Hoo, C. Brooks, D. Gunter. 7th IEEE Symposium on High Performance Distributed Computing, Chicago, Ill. July 29-31, 1998.
Netlogger provides services and tools for precision monitoring of application performance and communications, as well as related system and network events. http://www-didc.lbl.gov/NetLogger/

[27] WebFlow - A prototype visual graph based dataflow environment, WebFlow, uses the mesh of Java Web Servers as a control and coordination middleware, Rebuff. See http://iwt.npac.syr.edu/projects/webflow/index.htm

[28] A collaborative effort to enable desktop access to remote resources including, supercomputers, network of workstations, smart instruments, data resources, and more - computingportals.org

[29] ESNet - www.esnet.gov

[30] NREN - www.nren.nasa.gov